In Search of the Best Response Scale in a Mixed-mode Survey (Web and Mail): Evidence from MTMM Experiments in the GESIS Panel

Schwarz, Hannah; Weber, Wiebke; Minderop, Isabella; Weiß, Bernd

methods, data, analyses | 2021, Vol. 15(2), pp. 161-190 DOI: 10.12758/mda.2021.05

In Search of the Best Response Scale in a Mixed-mode Survey (Web and Mail). Evidence from MTMM Experiments in the GESIS Panel

Hannah Schwarz[1], *Wiebke Weber* [1], *Isabella Minderop*[2] *& Bernd Weiß*[2]

[1] *RECSM, Universitat Pompeu Fabra*

[2] *GESIS – Leibniz Institute for the Social Sciences*

Abstract

Mixed-mode surveys allow researchers to combine the advantages of multiple modes, for example, the low cost of the web mode with the higher coverage of offline modes. One drawback of combining modes is that there might be systematic differences in measurement across modes. Thus, it would be useful to know which measurement methods work best in all employed modes. This study sets out to find a method that results in the highest measurement quality across self-administered web mode questionnaires (web mode) and self-administered paper questionnaires sent out by mail (mail mode). Two Multitrait-Multimethod (MTMM) experiments employing questions on environmental attitudes and supernatural beliefs were implemented in the GESIS Panel, a probability-based panel in Germany. The experiments were designed to estimate the measurement quality of three different response scales: A seven-point fully labelled scale, a 101-point numerical open-ended scale and an eleven-point partially labelled scale. Our results show that the eleven-point partially labelled scale consistently leads to the highest measurement quality across both modes. We thus recommend using eleven-point partially labelled scales when measuring attitudes or beliefs in mixed-mode surveys combining web and mail mode.

Keywords: measurement quality; length of response scales; labelling of response scales; Multitrait-Multimethod (MTMM); mixed-mode; self-completion; web mode; mail mode

While in the past surveys were mainly unimode, nowadays respondents often receive the possibility to answer in the mode of their choice. This is supposed to increase their willingness to participate and lower survey costs (Eifler & Faulbaum, 2017). Mixed-mode surveys are used in various settings, especially where certain population groups are difficult to reach via the main survey mode. For example, these surveys are useful when researchers aim to conduct web mode surveys but have to account for the fact that parts of the target population do not use the internet (Bosnjak et al., 2017; ESOMAR & WAPOR, 2014). While this is an adequate strategy to deal with coverage error, it may lead to issues concerning measurement equivalence, as respondents may answer questions differently across modes (ESOMAR & WAPOR, 2014; Grewenig et al., 2018; Blom et al., 2016). Linked to this, employing the same measurement instrument across different modes can also lead to differences in measurement quality (e.g., Sánchez Tomé, 2018; Tourangeau, 2017; Dillman et al., 2014).

Measurement quality, in general, refers to the relationship between the unobserved, latent variable of interest and the observed response, and is here defined as the product of validity and reliability (Saris & Andrews, 1991). Validity here covers the construct validity subtypes convergent and discriminant validity (Campbell & Fiske, 1959). More specifically, it is defined as the strength of the relationship between a latent variable of interest and a so-called 'true score'. This 'true score' represents the score that respondents would have provided if no random measurement error existed. Reliability, in the model we employ, is defined as the strength of the relationship between the 'true score' and the observed variable. It captures the absence of random measurement error. It should be noted that an array of different definitions and operationalizations of measurement quality, as well as of validity and reliability, are used in the literature (Saris & Andrews, 1991). While studies may differ along these lines, they share the aim of empirically capturing measurement quality, that is, the absence of measurement error. The above definition of measurement quality thus holds for our analysis, while a somewhat broader perspective on the concept will be considered in the literature review.

Questionnaire designers in mixed-mode settings need not only to make sure that they indeed measure what they aim to measure. Furthermore, they need to ensure that respondents in different modes understand the measurement instru-

Acknowledgements

We thank Sonja Paulick-Fabini as well as the anonymous reviewer for their helpful comments.

Direct correspondence to

Hannah Schwarz, RECSM, Universitat Pompeu Fabra,
Sociometric Research Foundation
E-mail: hannah.schwarz@upf.edu

ment similarly. This means that questionnaire designers employing multiple modes have to make decisions bearing in mind the various features of the different survey modes. In self-administered modes, respondents can always see the questions and response scales, while in interviewer-administered modes they may only hear them. Furthermore, mail mode respondents see the questions on paper, while web mode respondents see them on electronic devices with varying screen sizes. Mail mode respondents answer using a pen or pencil while web mode respondents use a mouse, keyboard or touchscreen. Such features can have an influence on the comparability and quality of the measurement instrument.

Klausch et al.'s (2013) findings suggest that comparability of measurement between modes may not be attainable when comparing self-administered and interviewer-administered modes, but that measurement between different self-administered modes is comparable. Other authors have also reported this pattern (see e.g., De Leeuw & Hox, 2011; Hox et al., 2017). Yet, there are also studies finding differences within the group of self-administered modes, more specifically between mail and web mode, on aspects such as response quality, response patterns and estimation precision (Savage & Waldman, 2008; Olsen, 2009; Kwak & Radler, 2002). These differences are, for example, hypothesised to be due to online respondents suffering more fatigue and boredom which, in turn, could be caused by visual and interactive stimuli in the online mode being more cognitively demanding (Savage & Waldman, 2008). Kwak and Radler (2002) also discuss that differences in visual display, for example, different sizes of open-answer fields, or in the relative burden caused by filter questions in mail as compared to in web surveys, could cause such mode differences. Olsen (2009) attributes these mode differences to different self-selection processes into the mode groups.

If measurement differs between modes, different ways of designing a survey question, which we refer to as different methods in the following, might thus be preferable per mode to ensure the highest measurement quality. For example, longer response scales, i.e., with more answer categories (see e.g., Alwin, 1997; Andrews, 1984; Cox III, 1980; Költringer, 1995; Saris et al., 1977), as well as fully labelled response scales (see e.g., Alwin, 2007; Alwin & Krosnick, 1991; Saris & Gallhofer, 2007) tend to lead to higher measurement quality. However, one might not expect long and fully labelled scales to lead to high measurement quality in purely oral modes where respondents are unlikely to keep all response options equally present in their memory before answering (Krosnick & Alwin, 1987). Thus, long lists of response categories are typically not read out in oral modes (Schwarz et al., 1991). For unimode surveys, method recommendations tailored to the employed mode may thus be followed. However, where comparability across modes is crucial, such as in mixed-mode surveys, the focus should be on finding those methods that lead to the highest measurement quality in all modes used.

Various question characteristics have been studied in terms of their links to measurement quality (see e.g., DeCastellarnau, 2018). While, in practice, question characteristics are often interrelated and there are no incontestable unique guidelines on what works best (DeCastellarnau, 2018; Saris & Gallhofer, 2014; Schaeffer & Dykema, 2020), this research helps questionnaire designers. It enables them to carefully consider the way they employ question characteristics in their measurement instruments, taking into account different theoretical arguments and empirical evidence. Previous research has determined the measurement quality of specific questions (see e.g., Revilla et al., 2014; Oberski et al., 2007) as well as the influence of question characteristics on measurement quality through meta-analysis (Kogovšek & Ferligoj, 2005; Saris & Gallhofer; 2014; Saris et al., 2011; Scherpenzeel & Saris, 1997). Such research has been conducted in several countries, concerning various question topics and in different modes of data collection. Still, as web surveys have existed for a relatively short time, measurement quality assessments for this mode are still rarer than for other modes (Bosch et al., 2019). Furthermore, web panels are a special context in which web surveys are administered, on which even less research exists. The specificity here comes particularly from the fact that panel conditioning, i.e., training or learning effects, can appear, which tend to lead to an increase in the reliability and stability of responses over time (Sturgis et al., 2009). Moreover, as in most countries substantive parts of the population do not use the internet (World Bank, 2020), it is crucial to also study the measurement quality of survey questions in mixed-mode settings (Callegaro et al., 2014).

This study therefore sets out to assess measurement quality in a mixed-mode panel survey, using web and mail mode, to find a measurement method that results in the highest measurement quality across both modes. We do this by conducting two Multitrait-Multimethod (MTMM) experiments, allowing us to estimate measurement quality as defined above. Furthermore, to advance research on the links between question characteristics and measurement quality, we particularly focus on the effect of two response scale characteristics, namely the length and labelling of response scales.

This paper proceeds as follows: We first present the theoretical argumentation and empirical evidence concerning using response scales of a certain length and using fully versus partially labelled response scales. On this basis, we formulate hypotheses. We then describe the data, the experimental design and the analytical strategy. Subsequently, we present the results, discuss them and draw conclusions.

Theory and Empirical Evidence: Scale Characteristics and their Impact on Measurement Quality

Length of Response Scales: Theory

Much of the literature on the relationship between response scale length and measurement quality bases its theoretical argument on the theory of information (e.g., Alwin, 2007; Alwin et al., 2018; Revilla et al., 2014). The theory of information suggests that with an increasing number of scale points, not only the direction but also the intensity or extremity of an attitude can be assessed in an increasingly detailed fashion (Garner, 1960). Therefore, longer scales should result in better measurement quality because more information can be gathered (see also Alwin, 1997; Andrews, 1984; Cox III, 1980; Költringer, 1995; Saris et al., 1977). Along the same lines, Alwin and Krosnick (1991) describe that offering too few categories would lead to a loss of information, as respondents would have to 'round' their answers.

However, there are also arguments for not including too many answer categories. Schaeffer and Presser (2003) state that the right response scale length should be a compromise between offering more potential for finer distinctions and considering respondents' limited capacities for making finer distinctions reliably and in similar ways. For example, a 100-point scale enables respondents to make finer distinctions than a five-point scale. However, it bears higher potential to induce different responses from a respondent when asked repeatedly across time, as well as to be used in different ways across respondents compared to a five-points scale. Similarly, other authors argue against the use of long response scales, referring to the suggestion of cognitive theorists that there is an upper limit to how many answer categories respondents can handle (Vall-Llosera et al., 2020). At a certain point, adding more categories results in the answer options having less rather than more meaning. Moreover, referring to motivational theories, the task of answering survey questions becomes increasingly complex the more answer categories are offered, thus too many scale points could lead to satisficing (Alwin, 1997). Especially scholars discussing very long scales have pointed out that respondents are likely to engage in rounding which can be regarded as a form of satisficing because, rather than considering all answer options, the task complexity is reduced by effectively only considering a part of the answer options (Liu & Conrad, 2016; Tourangeau et al., 2000).

Length of Response Scales: Empirical Evidence

Previous studies offer a large body of empirical findings on the optimal length of response scales. It should be noted that response scales can be classified in terms of various characteristics, such as the scales' evaluative dimension (item-specific versus agree-disagree) or the scales' polarity (unipolar versus bipolar) (DeCastellarnau, 2018). These characteristics are interrelated. For example, agree-disagree scales are always bipolar. To review the empirical literature on the impact of single response scale characteristics on measurement quality, we will draw from findings on scales that are otherwise heterogenous. For example, we will look at the effects of response scale length in both item-specific and agree-disagree scales, to gather as many findings as possible on the impact of response scale length on measurement quality. Moreover, due to the mixed-mode angle of this study, we will also consider whether results differ across modes in our review. Where operationalizations of measurement quality, reliability or validity diverge from the operationalization we use, this will be indicated in the following by specifying the exact indicator used (e.g., test-retest reliability) or by describing the operationalization.

Many scholars report an improvement of measurement quality with an increase of answer categories. For example, Alwin (1997) finds higher reliability and validity for eleven-point scales than for seven-point scales in a study employing face-to-face mode. Andrews (1984) also finds that using more categories increases measurement quality, both in terms of reliability and validity, in a study conducted in various modes, namely telephone, face-to-face and group interviews. Rodgers et al. (1992) find in a face-to-face study that both validity and reliability increase with the number of scale points. Lundmark et al. (2016) look at concurrent validity, i.e., the extent to which a variable can predict other variables it should be related to. They find this to be higher for longer scales (seven and eleven-point scales as compared to two-point scales) in a web mode survey. Furthermore, Wu and Leung (2017) use simulated survey data to compare scales of four, five, six, seven and eleven points and find the longer scales to lead to higher measurement quality, here defined as the accordance of the simulated data with the 'true scores' calculated from a known underlying distribution. Revilla and Ochoa (2015) similarly find longer scales to lead to better measurement quality, at least up to eleven points, focusing on item specific scales in a web survey. Yet, looking specifically at agree-disagree scales, Revilla et al. (2014) do not find measurement quality to improve by increasing the number of scale points beyond five. Their results are based on a face-to-face study.

Many authors find that improving measurement quality by increasing the number of answer categories only works up to a certain point beyond which no improvements are observed. Instead, quality might even decrease. This is often described as a curvilinear effect. For example, Preston and Colman's (2000) findings suggest a curvilinear effect when looking at test-retest reliability: Adding cat-

egories increases this measure of reliability between two and ten scale points, but adding further points leads to a decrease in test-retest reliability. They find a similar pattern when looking at indicators of criterion and convergent validity. Their study was conducted using self-administered paper questionnaires. Similarly, Saris and Gallhofer (2007) find that increasing the number of categories up to eleven points leads to improved measurement quality in their meta-analysis based on data from face-to-face interviews, the disk-by-mail approach[1] and the Telepanel[2]. Alwin and Krosnick (1991), using data from face-to-face interviews, find that for item specific questions, the quasi-simplex model reliability[3] increases from three to seven points and then remains constant when the scale is extended to nine points.

In contrast, other scholars find relatively short scales to be superior. McKelvie (1978) finds that test-retest reliability tends to be highest when using five-point scales in his study using self-administered paper questionnaires. Alwin (2007), looking particularly at unipolar scales and using quasi-simplex models, finds that they are most reliable at four points. He bases his findings on a mix of face-to-face and self-administered paper questionnaire surveys. Scherpenzeel and Saris (1997) stress the different effects response scale length can have on validity and reliability showing that validity is highest at four, five or seven points while reliability is highest at two to three points. They analyse data from web surveys, mail surveys and computer assisted telephone interviews (CATI). They also look at potential differences between modes but do not find any. Alwin et al. (2018) find that reliability tends to decline with an increasing number of response options, with two-point scales resulting in the highest reliability. Unipolar measures of attitudes form the exception. For this type of question, reliability increases with longer scales. Their analysis is based on General Social Survey questions conducted in face-to-face mode.

There are also studies suggesting that changing the number of response categories does not affect measurement quality. Jacoby and Matell (1971), looking at both test-retest reliability and indicators for predictive and concurrent validity, find this in their study focusing on agree-disagree scales based on self-administered paper questionnaires. McKelvie (1978) also finds indications for this, at least in terms of validity, in his study using self-administered paper questionnaires. More precisely, he does not find criterion validity, operationalized as correlating responses with

1 A floppy disk containing the survey and the programme required to open the survey was sent to respondents.

2 An early web mode approach. Respondents were provided with a computer and a modem, if necessary, so that surveys could be sent to them.

3 The quasi-simplex model is an extension of the test-retest model using at least three repeated measures of the same variable over time to estimate reliability. It allows to account for change in the measure of interest and assumes that there is no method effect (Saris & Gallhofer, 2014).

available objectively correct values, to be affected by a change in the number of answer categories.

Overall, most empirical findings seem to suggest that longer scales can indeed lead to higher measurement quality but that this only works up to a certain point from which on quality tends to remain stable. Yet, different studies find different optimal numbers of scale points, ranging from five to eleven. From this review, we cannot deduce that this should differ between web and mail mode. We therefore expect that *response scales with five to eleven points result in the highest measurement quality in both web and mail mode (H1).*

Fully Labelled Versus Partially Labelled Response Scales: Theory

More comprehensive labelling of a scale is commonly assumed to be beneficial as it clarifies the meaning of otherwise ambiguous scale points, thus reducing variability in scale point interpretation across respondents (Alwin, 2007; Eutsler & Lang, 2015; Krosnick & Berent, 1993; Krosnick & Fabrigar, 1997). Verbal labels should be a more natural form of expressing meaning compared to numbers (Krosnick & Fabrigar, 1997). Receiving the information in text form, rather than via numbers, should therefore reduce respondent burden (Krosnick & Presser, 2010).

Yet, there are arguments that suggest more extensive verbal labelling might be harmful to measurement quality. For example, Krosnick and Fabrigar (1997) mention that verbal labels could be problematic due to language ambiguity and are also more difficult to remember (see also Alwin & Krosnick, 1991). They argue that the task of answering a survey question could be less cognitively demanding for respondents if they have to read fewer labels, for example, when only end point labels are used (see also Kunz, 2015). This stands in direct contrast with the argument made above. Menold et al. (2014) reconcile these opposing assumptions stating that while full verbal labelling facilitates interpretation, it makes the mapping process more burdensome when compared to end point labelling.

Fully Labelled Versus Partially Lablled Response Scales: Empirical Evidence

The vast majority of research finds fully labelled scales to be superior to partially labelled ones of similar length. For example, Alwin (2007) finds the quasi-simplex model reliability of response scales to increase when full labels rather than just endpoint labels are used. He bases his work on a variety of face-to-face and self-administered paper questionnaire surveys (administered on site, i.e., not mail mode). Alwin and Krosnick (1991) find that using fully labelled response options is

associated with an increase in quasi-simplex model reliability in a study based on face-to-face and telephone surveys. Similarly, Saris and Gallhofer (2007), in their study based on data from face-to-face interviews, the disk-by-mail approach and the Telepanel find that the use of verbal labels increases the reliability of questions.

There are, however, also some findings that point in the opposite direction. Andrews (1984) concludes from his analyses of data collected in telephone and face-to-face individual and group interviews that measurement quality decreases where fully labelled answer categories are used. Similarly, Rodgers et al. (1992) find full labelling to lead to more random measurement error, i.e., lower reliability, in a face-to-face survey.

To sum up, most empirical assessments of the issue find that fully labelled scales lead to higher measurement quality. This was found to be the case across various modes. We therefore expect that *fully labelled response scales lead to higher measurement quality in both web and mail mode* (H2).

Comparing the Effects of Scale Length and Full Labelling

So far, we have focused on the impact of the length of response scales and the labelling of response scales separately. However, for the sake of deriving practical recommendations for questionnaire designers, we would also like to assess whether it is more beneficial for measurement quality to have a long or a fully labelled response scale. We could only find one study that compared the effect of these two characteristics on measurement quality based on a meta-analysis. Andrews (1984) shows that the number of scale categories explains a larger share of the variance in validity and reliability than the labelling of the scale. Therefore, we expect that *the benefit of employing longer response scales will outweigh the benefit of employing fully labelled response scales* (H3).

Data and Method

Sample

We conduct the experiments in the GESIS Panel, a probability-based mixed-mode panel in which about 75% of the respondents answer in web mode and 25% in mail mode. The GESIS Panel was founded in 2013 and contains about 5000 panelists. To account for attrition, the sample was refreshed in 2016 and 2018. Every two months, panelists are invited to participate in a survey lasting approximately 20 minutes. They receive a five-euro prepaid incentive with each survey invitation (GESIS, 2020; Minderop et al., 2019; Bosnjak et al., 2017). Upon a face-to-face recruitment interview, those respondents who indicated that they use the internet regularly were

Table 1 Characteristics of sample before listwise deletion for both web and mail mode (unweighted)

	Web mode			Mail mode		
	Mean	SD	Valid n	Mean	SD	Valid n
Age	47.08	14.36	2,779	57.19	12.69	1,028
Female	49.14%	.50	2,784	54.07%	49.86	1,032
University education	34.18%	47.44	2.762	13.76%	34.46	1,025
Total			2,784			1,032

offered to participate in web mode. Interviewers were requested to present online participation as an attractive option and to persuade respondents to participate in web mode. However, internet users were also free to opt for mail mode. Those respondents who did not use the internet were only presented the option to participate in mail mode (Bosnjak et al., 2017). The Multitrait-Multimethod experiments were implemented in the 'gb' wave fielded in April and May 2019 (GESIS, 2020).

Sociodemographic characteristics of both web mode and mail mode respondents in the sample before listwise deletion[4] of respondents with missing values on the experimental variables are presented in Table 1. As can be expected, respondents who self-selected into the web mode differ significantly from those who self-selected into the mail mode. Mail respondents are on average about ten years older than web respondents ($p<.001$). Furthermore, the proportion of female respondents is about five percentage points higher among mail respondents than among web respondents ($p<.05$). Women are thus overrepresented in mail mode. The proportion of respondents who have obtained a university degree is substantially higher among web mode respondents (34%) than among mail mode respondents (14%) ($p<.001$). After listwise deletion of cases with missing values, the total valid sample size for experiment 1 (environmental attitudes) is n=3,632 and n=3,589 for experiment 2 (supernatural beliefs). We also conduct analyses of variance to check if sociodemographic characteristics differ significantly across the experimental groups. The results show that differences approach significance ($p=.0596$) only for 'university education'. Concretely, the proportion of respondents who indicated that a university degree is their highest achieved level of education is about four percentage points lower for group two (26.34%) than for groups one and three (30.29% and 29.34%, respectively). As this difference is substantively small, we see no rea-

4 We ran a robustness analysis using pairwise deletion instead. The resulting estimates are extremely similar to those found using listwise deletion. The results would not lead to an alteration of any substantive findings.

son to be concerned about the success of respondents' random assignment into experimental groups.

The True Score MTMM Model

The MTMM experimental design used here is based on the True Score MTMM (TS-MTMM) model proposed by Saris and Andrews (1991) to estimate the reliability, validity, and quality of the survey questions. According to Saris and Andrews (1991), measurement quality is defined as the product of validity and reliability. Validity is defined as the strength of the relationship between a latent variable of interest and the 'true score' and reliability as the strength of the relationship between the 'true score' and the observed variable.

The following system of equations describes the TS-MTMM model:

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \tag{1}$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \tag{2}$$

with F_i being the i^{th} trait or factor, M_j being the j^{th} method, Y_{ij} being the observed answer for the i^{th} trait and the j^{th} method, T_{ij} being the true score factor or systematic component of the response, r_{ij} being the reliability coefficient (when standardized), v_{ij} being the validity coefficient (when standardized), and e_{ij} being the random error associated with Y_{ij}.

Equation (1) defines the observed variables as the sum of the associated systematic component and random errors. Equation (2) defines the systematic components themselves as the sum of the trait component and the effect of the method employed to assess the trait. The total measurement quality can be obtained by taking the product of the reliability and validity, being the reliability coefficient and the validity coefficient squared: An illustration of the path diagram of the True Score MTMM model for three traits, each measured with three methods is presented in Figure 1.

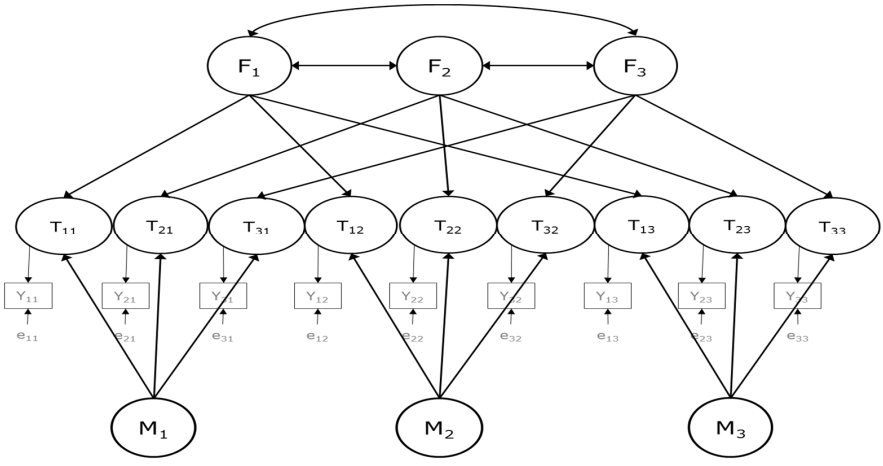


Figure 1 Path diagram of the True Score MTMM model for three traits and three methods

It shows that each trait (F_i) is measured three times with different methods (M_j). This results in nine true scores (T_{ij}) which are measured by the nine survey questions that are evaluated in each experiment. The observed responses to these nine questions are denoted as Y_{ij}. By measuring three correlated traits with three methods, we can thus estimate the measurement quality of all employed survey questions estimating the TS-MTMM model using Structural Equation Modelling (SEM) (see also section on analytical strategy).

As Figure 1 shows, we assume the traits (F_i) to be correlated, the method factors (M_j) to be uncorrelated, and the method factors to be uncorrelated with the trait factors. We also assume that the impact of the method factor on the traits measured with a common scale is the same and that the random errors (e_{ij}) are uncorrelated with each other and with the true scores (T_{ij}), the trait factors (F_i) and the method factors (M_j).

The Assessed Traits

The traits for experiment 1 are three questions on environmental attitudes based on previous questions asked on the GESIS Panel (GESIS, 2020). The traits for experiment 2 are three questions in supernatural beliefs based on questions from the ALLBUS 2012 (GESIS, 2016). Table 2 shows English translations of the questions. The German questions can be found in Appendix B.

Table 2 Traits

Experiment 1: Environmental attitudes	
Trait 1	Can you identify with environmentalists?
Trait 2	Should we all be willing to restrict our current living standard for the benefit of the environment?
Trait 3	Do you believe that some problems of our times would be solved if we went back to a more rural and natural lifestyle?
Experiment 2: Supernatural beliefs	
	How much do you believe in the following?
Trait 1	…in life after death
Trait 2	…in heaven
Trait 3	…in miracles

The Assessed Methods

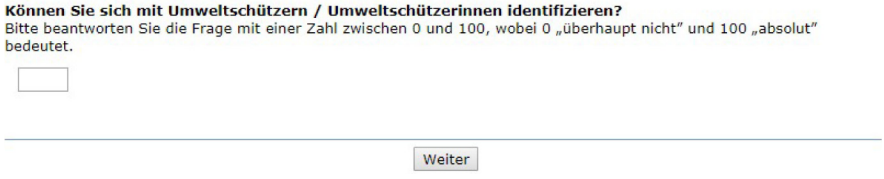
To test our hypotheses, we focus on varying the length of the response scales and the extent of labelling answer categories. However, to be able to identify the MTMM model in the analysis, it is helpful to vary further question characteristics. In the three assessed methods, we vary the following characteristics (see also Table 3): (1) the length of the response scale; (2) the verbal labelling of the response scale (fully versus partially labelled); (3) whether a continuous or discrete scale is used; (4) whether the scale is presented in a horizontal format or as a numerical open-ended scale; (5) whether a definition of the scale is present in the request or not. Figures 2 and 3 display how the methods for the first trait of the first experiment appear in the GESIS Panel web and mail questionnaire, respectively. In Appendix A, we present an exemplary smartphone screenshot, showing that the horizontal response scales were also displayed horizontally on small screen mobile devices.

Table 3 Variations of question characteristics across methods

Variation	Method 1	Method 2	Method 3
1	7-points	101-points	11-points
2	Fully labelled	Partially labelled	Partially labelled
3	Discrete	Continuous	Discrete
4	No definition of scale	Definition of scale	Definition of scale
5	Horizontal	Numerical open-ended scale	Horizontal

Method 1:

Method 2:



Method 3:

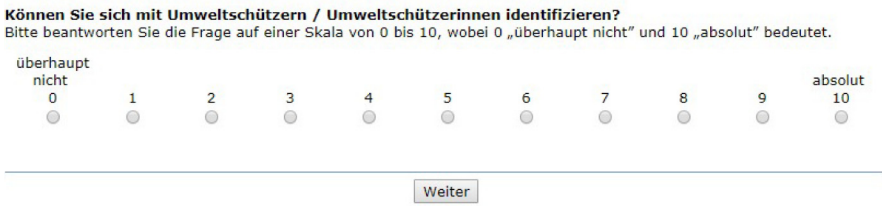


Figure 2 Screenshots of the GESIS Panel web questionnaire: Trait 1 of experiment 1 asked with methods 1, 2 and 3

Method 1:

(50) Können Sie sich mit Umweltschützern / Umweltschützerinnen identifizieren?

überhaupt nicht	ein wenig	etwas	mäßig	erheblich	sehr	absolut
o	o	o	o	o	o	o

Method 2:

(50) Können Sie sich mit Umweltschützern / Umweltschützerinnen identifizieren?

Bitte beantworten Sie die Frage mit einer Zahl zwischen 0 und 100, wobei 0 „überhaupt nicht" und 100 „absolut" bedeutet.

___

Method 3:

(50) Können Sie sich mit Umweltschützern / Umweltschützerinnen identifizieren?

Bitte beantworten Sie die Frage auf einer Skala von 0 bis 10, wobei 0 „überhaupt nicht" und 10 „absolut" bedeutet.

überhaupt nicht										absolut
0	1	2	3	4	5	6	7	8	9	10
o	o	o	o	o	o	o	o	o	o	o

Figure 3 Depiction of the GESIS Panel mail questionnaire: Trait 1 of experiment 1 asked with methods 1, 2 and 3

Experimental Design

For the experiments, respondents are randomly assigned to three groups of approximately equal size. In three-group Split Ballot MTMM experiments, each group receives three questions asking for the three traits using one method at time 1 (towards the middle of the questionnaire), and each group receives the same three questions again but with another method at time 2 (at the end of the questionnaire). Other questions are asked in between the two instances to reduce memory effects (Schwarz et al., 2020; Van Meurs & Saris, 1990). By implementing two methods in each group but varying which methods these are across groups, all combinations of methods are covered. Also, respondents do not have to handle the burden and potential fatigue that would result from asking them the same questions three times. As we run two MTMM experiments in one survey, we vary which groups are asked with which methods across the two experiments (see Table 4) to avoid repetitions of the same methods within a group as much as possible.

Table 4 Three-group Split-Ballot MTMM Design for both experiments

	Time 1		Time 2	
	Experiment 1	Experiment 2	Experiment 1	Experiment 2
Group 1	M1	M2	M2	M3
Group 2	M2	M3	M3	M1
Group 3	M3	M1	M1	M2

Analytical Strategy: Model Estimation and Testing

For model estimation we use Maximum Likelihood in LISREL 8.72 (Jöreskog & Sörbom, 1996). The base model in LISREL notation can be found in Appendix C. We run three separate analyses: (1) for the entire sample, (2) only for web mode respondents and (3) only for mail mode respondents. For testing, we evaluate the local model fit with the software JRule (Van der Veld et al., 2008). Parameter misspecifications indicated by JRule are used to improve the model. Such improvements can consist in allowing unequal effects of one method on the different traits, freeing error variances because of timing effects, adding a correlation between two methods, or allowing correlations between errors due to expected memory effects. As we expect the same models to hold in the analysis of the entire sample as well as presenting reliability and validity separately, we aim to implement the same adjustments to the model across these analyses. However, this is not always possible (i.e., it can result in improper solutions or poor model fit). The final model adjustments for all analyses are shown in Appendix D, as are the global model fit indices and indications of remaining local misspecifications as shown by JRule.

Results

In Table 5, we present the average measurement quality across traits by method, experiment and mode. The detailed results, i.e., per trait as well as presenting reliability and validity separately, are shown in Appendix E. We consider a quality estimate above or equal to .9 to indicate excellent measurement quality, a quality estimate between .9 and .8 good quality and a quality estimate between .8 and .7 acceptable quality. A quality estimate between .7 and .6 is seen as questionable, and quality estimates below .6 are interpreted as poor measurement quality.

Table 5 Average quality across all traits by method, mode of data collection and experiment

	Experiment: Environmental Attitudes			Experiment: Supernatural Beliefs			Both experiments		
	Both modes	Web	Mail	Both modes	Web	Mail	Both modes	Web	Mail
M1	0.69	0.73	0.59	0.90	0.89	0.89	0.80	0.81	0.74
M2	0.72	0.77	0.66	0.89	0.90	0.86	0.80	0.83	0.76
M3	0.83	0.83	0.79	0.99	1.00	0.97	0.91	0.91	0.88

Note: M= Method; M1= 7-point fully labelled horizontal, no scale definition; M2= 101-point numerical open-ended scale, scale definition present; M3= 11-point partially labelled horizontal, scale definition present.

Looking at all average quality estimates across methods and experiments, we find that measurement quality tends to be especially high for the experiment on supernatural beliefs, with all estimates indicating excellent or good quality. Findings are more mixed for the experiment on environmental attitudes, with quality estimates ranging from good to questionable and even to poor in one instance (for method 1 in mail mode).

When we look at the average quality for each method in both experiments, we find that, overall, method 3 (eleven points, only end points labelled) obtains the highest quality (between .79 and 1), independently of the mode of data collection or the topic of the experiment. Thus, our hypothesis that the benefits of using long response scales outweigh the benefits of using fully labelled response scales (H3) cannot be rejected. Comparing the performance of method 1 (seven-point fully labelled) and method 2 (numerical open-ended scale ranging from zero to 100) in all modes and experiments shows that they perform similarly. An exception can be observed in mail mode in experiment 1, where method 1 performs substantially worse than method 2. The similar performance of methods 1 and 2 is not in line with our expectation formulated in H1 that scales between five and eleven points result in the highest measurement quality. Instead, our results show that the 101-point scale tends to result in the same measurement quality as the seven-point scale. For the experiment on environmental attitudes, it appears that the longer scale even outperforms the seven-point scale, at least for mail mode. Moreover, the observation that method 1 results in the lowest measurement quality in most instances and that it is consistently outperformed by the partially labelled scale (method 3) means we can reject H2 that fully labelled response scales lead to higher measurement quality.

Furthermore, the observation that method 3 performs best across all modes and that there are few differences in the performance of methods 1 and 2 in the different modes also means that it is indeed possible to find one method that performs best across both modes in this case[5].

Discussion and Conclusion

In this paper we set out to assess which response scale results in the highest measurement quality across two modes of data collection, self-completion in a web survey (web mode) and on a paper questionnaire (mail mode). Given the differing internet penetration and internet literacy across and even within countries, such mixed-mode designs are a valuable option to increase survey participation while saving costs. Based on the state-of-the-art in the field, we formulated hypotheses regarding the impact of length and labelling of response scales on measurement quality.

In line with the literature, we find that the eleven-point partially labelled scale (method 3) consistently produces the highest measurement quality across modes for both experiments (Preston & Colman, 2000; Rodgers et al., 1992; Saris & Gallhofer, 2007). Contrary to previous results, we find that a numerical open-ended scale, i.e., a scale requiring respondents to indicate the answer using a number, here between zero and 100, and a seven-point fully labelled response scale tend to result in the same measurement quality. Previous literature has found fully labelled scales to lead to higher measurement quality across various modes including self-completion on the web (Saris & Gallhofer, 2007) and on paper questionnaires (Alwin, 2007).

Moreover, differences in measurement quality across modes have been reported (Sánchez Tomé, 2018; Tourangeau, 2017; Dillman et al., 2014). However, our study suggests that there are no systematic differences across modes concerning the effect of response scale length and labelling on measurement quality.

Furthermore, we find that using longer response scales seems to give more of a boost to measurement quality than using fully labelled scales (H3). The partially labelled eleven-point scale (method 3) outperforms the fully labelled seven-point scale (method 1) consistently, and the numerical open-ended scale (method 2) outperforms the fully labelled seven-point scale (method 1) for one mode in one experiment. However, longer scales are not generally better. Our study shows that

5 We also ran a robustness analysis on only respondents using smartphones (valid n for the experiment on environmental attitudes is 512 and for the experiment on supernatural beliefs is 516). The resulting estimates are extremely similar to those found for web mode overall. Analysing smartphone respondents separately leads to the same substantive findings.

increasing the number of scale points from seven to eleven yields higher measurement quality but increasing it from eleven to 101 points leads to inferior measurement quality.

On the basis of these findings, we can recommend using an eleven-point partially labelled scale (method 3) when measuring attitudes or beliefs in mixed-mode surveys combining web and mail mode. Furthermore, we recommend prioritizing the use of longer response scales (up to eleven points) over the use of seven-point fully labelled scales.

One limitation of our study results from the suboptimal formulation of the questions of experiment 1. Question formulations here did not indicate that respondents would be able to give a nuanced answer but read as yes/no questions. This might partly explain why lower measurement quality is obtained by the questions of experiment 1 compared to those of experiment 2.

Another limitation is that, given the design of our experiments, we cannot draw conclusions beyond the particular combinations of characteristics present in the tested response scales. To estimate an MTMM model, several scale characteristics should be varied across methods. Therefore, we could not assess the isolated effect of one scale characteristic. To do so, more experiments and a meta-analysis are needed (Kogovšek & Ferligoj, 2005; Saris & Gallhofer, 2014; Saris et al., 2011; Scherpenzeel & Saris, 1997). However, in terms of practical implications it is not always necessary to unconfound the impact of different question characteristics. In survey practice, specific question characteristics tend to occur together (e.g., eleven-point scales are usually only partially labelled), while the combination of other question characteristics is less practically feasible, less common, and therefore less relevant to study (e.g., eleven-point scales are rarely fully labelled). These "structural dependencies among sets of characteristics" are also pointed out by Schaeffer and Dykema (2020, p.10.6), reminding us that decisions in the design of survey questions depend on what combinations of characteristics can or cannot occur together. The results of MTMM experiments showing which measurement scales lead to which measurement quality thus remain a vital basis for questionnaire design.

Further research is needed to investigate the questions left open by this study: Does full labelling only lead to higher measurement quality in shorter scales? Are eleven points really the optimal length, or may slightly shorter scales (for example: nine points) be preferable? If the seven-point scale had been only partially labelled, would it still be outperformed by the partially labelled eleven-point scale? And would any of these adjustments have resulted in differences across modes? In short, a variety of feasible scales remain to be tested on mixed-mode panels such as the GESIS Panel and mode differences should always be taken into account.

Bibliography

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research*, *25*(3), 318-340. doi:10.1177/0049124197025003003

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods and Research*, *20*(1), 139–181. doi:10.1177/0049124191020001005

Alwin, D. F., Baumgartner, E. M., & Beattie, B. A. (2018). Number of response categories and reliability in attitude measurement. *Journal of Survey Statistics and Methodology*, *6*(2), 212-239. doi:10.1093/jssam/smx025

Alwin, D.F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken: Wiley.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409-442. doi:10.1086/268840

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A. S., Das, M., Douhou, S., & Krieger, U. (2016). A Comparison of four probability-based online and mixed-mode Panels in Europe. *Social Science Computer Review, 34*(1), 8-25. doi:10.1177/0894439315574825

Bosch, O. J., Revilla, M., DeCastellarnau, A., & Weber, W. (2019). Measurement reliability, validity, and quality of slider versus radio button scales in an online probability-based panel in Norway. *Social Science Computer Review, 37*(1), 119-132. doi:10.1177/0894439317750089

Bosnjak, M., Dannwolf, T., Enderle, T., Schauer, I., Struminskaya, B., Tanner, A., & Weyandt K. W. (2017). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review, 36*(1). doi:10.1177/0894439317697949

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 1-22). Hoboken: Wiley.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81-105.

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407-422. doi:10.1177/002224378001700401

De Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (pp. 45-76). New York: Routledge.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523-1559. doi:10.1007/s11135-017-0533-4

Dillman, D. A., Smyth, J. D. & Christian, L. M. (2014) Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. Hoboken: Wiley.

Eifler, S., & Faulbaum, F. (2017). Vorwort. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 7-8). Wiesbaden: Springer VS.

ESOMAR & WAPOR (2014). ESOMAR/WAPOR Guideline on Opinion Polls and Published Surveys (World Research Codes and Guidelines). Retrieved from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-WAPOR-Guideline-on-Opinion-Polls-and-Published-Surveys-August-2014.pdf

Eutsler, J., & Lang, B. (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting, 27*(2), 35-51. doi:10.2308/bria-51219

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67*(6), 343-352. doi:10.1037/h0043047

GESIS (2016). *ALLBUS 1980-2014 – Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ZA4582 Datenfile Version 1.0.0*, [Data file] Gesis – Leibniz Institute for the Social Sciences. doi:10.4232/1.12439

GESIS (2020): *GESIS Panel - Standard Edition. GESIS Data Archive, Cologne. ZA5665 Datenfile Version 33.0.0*, [Data file] Gesis – Leibniz Institute for the Social Sciences. doi:10.4232/1.13377

Grewenig, E., Lergetporer, P., Simon, L., Werner, K., & Woessmann, L. (2018). *Can online surveys represent the entire population?* (CESifo Working Paper No. 7222). Retrieved from https://ssrn.com/abstract=3275396

Hox, J. de Leeuw, E. & Klausch, T. (2017) Mixed mode research: Issues in design and analysis. In Biemer, P., de Leeuw, E. Eckman, S., Edwards, B., Kreuter, F.,Lyberg, L. E., Tucker, N. C. & West, B.T. (Eds.), *Total Survey Error in Practice* (pp. 511-530). Hoboken: Wiley.

Jacoby, J., & Matell, M. S. (1971). Three-point Likert Scales are good enough. *Journal of Marketing Research*, *8*(4), 495–500. doi:10.2307/3150242

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Uppsala, Sweden: Scientific Software International.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research, 42*(3), 227-263. doi:10.1177/0049124113500480

Kogovšek, T., & Ferligoj, A. (2005). The quality of measurement of personal support subnetworks. *Quality and Quantity, 38*(5), 517-532. doi:10.1007/s11135-005-2178-y

Költringer, R. (1995). Measurement quality in Austrian personal interview surveys. In W. E. Saris & A. Münnich (Eds.), *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments* (pp. 207–224). Budapest: Eötvös University Press.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identifications and policy preferences: the impact of survey question format. *American Journal of Political Science*, *37*(3), 941–964. doi:10.2307/2111580

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141-164). Hoboken: Wiley.

Krosnick, J.A., & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden, & J.D. Write (Eds.), *Handbook of Survey Research* (pp. 263–313). Bingley: Emerald.

Kunz, T. (2015). *Rating scales in web surveys. A test of new drag-and-drop rating procedures* (Doctoral dissertation). Retrieved from https://tuprints.ulb.tu-darmstadt.de/5151/7/Kunz_2015_Rating_scales_in_web_surveys.pdf

Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics, 18*(2), 257.

Liu, M., & Conrad, F. G. (2016). An experiment testing six formats of 101-point rating scales. *Computers in Human Behavior, 55*, 364-371. doi:10.1016/j.chb.2015.09.036

Lundmark, S., Gilljam, M., & Dahlberg, S. (2016). Measuring generalized trust. An examination of question wording and the number of scale points. *Public Opinion Quarterly*, *80*(1), 26–43. doi:10.1093/poq/nfv042

McKelvie, S. J. (1978). Graphic rating scales - How many categories? *British Journal of Psychology, 69*(2), 185-202. doi:10.1111/j.2044-8295.1978.tb01647.x

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*(1), 21-39. doi:10.1177/1525822X13508270

Minderop, I., Bretschi, D., Kolb, J., & Heycke, T. (2019). *GESIS Panel Wave Report Wave gb* (April/May 2019). Gesis – Leibniz Institute for the Social Sciences.

Oberski, D., Saris, W. E., & Hagenaars, J. (2007). Why are there differences in measurement quality across countries. In G. Loosveldt, M. Swyngedouw & B. Cambré (Eds.), *Measuring Meaningful Data in Social Research* (pp. 281-300). Leuven: Acco.

Olsen, S. B. (2009). Choosing between internet and mail survey modes for choice experiment surveys considering non-market goods. *Environmental and Resource Economics, 44*(4), 591-610. doi: 10.1007/s10640-009-9303-7

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15. doi:10.1016/S0001-6918(99)00050-5

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods and Research, 43*(1), 73-97. doi:10.1177/0049124113509605

Revilla, M., & Ochoa, C. (2015). Quality of different scales in an online survey in Mexico and Colombia. *Journal of Politics in Latin America*, *7*(3), 157–177. doi:10.1177/1866802X1500700305

Rodgers, W. L., Andrews, F. M., & Herzog, A. R. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, *8*(3), 251–275.

Sánchez Tomé, R. (2018). The impact of mode of data collection on measures of subjective wellbeing (Doctoral dissertation). Retrieved from https://serval.unil.ch/resource/serval:BIB_F89D8660FBE7.P001/REF

Saris W. E., Bruinsma, C., Schoots, W., & Vermeulen, C. (1977). The use of magnitude estimation in large scale survey research. *Mens en Maatschappij, 52* (4), 369–395.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P., Groves, R. E., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.), *Measurement Error in Surveys* (pp. 575-597). Hoboken: Wiley.

Saris, W. E., & Gallhofer, I. (2007) Design, Evaluation, and Analysis of Questionnaires for Survey Research. Hoboken: Wiley.

Saris, W. E., & Gallhofer, I. (2014) Design, Evaluation, and Analysis of Questionnaires for Survey Research. Hoboken: Wiley.

Saris, W., Oberski, D., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I., & Gruner, T. (2011). *The development of the Pprogram SQP 2.0 for the prediction of the quality of survey questions* (RECSM Working Paper No. 24). Retrieved from https://www.upf.edu/documents/3966940/3986764/RECSM_wp024.pdf

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics, 23*(3), 351-371. doi:10.1002/jae.984

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology, 46,* 37-60. doi:10.1146/annurev-soc-121919-054544

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65-88. doi:10.1146/annurev.soc.29.110702.110112

Scherpenzeel, A.C., & Saris, W.E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods Research, 25*(3), 341–383. doi:10.1177/0049124197025003004

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory effects in repeated survey questions: Reviving the empirical investigation of the independent measurements assumption. *Survey Research Methods, 14*(3). doi:10.18148/srm/2020.v14i3.7579

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*(3), 193-212.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In Lynn, P. (Ed.), *Methodology of longitudinal surveys* (pp. 113-126). Hoboken: Wiley.

Tourangeau, R. (2017). Mixing modes. In Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C. & West, B.T. (Eds.), *Total Survey Error in Practice* (pp. 115-132). Hoboken: Wiley.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press.

Vall-Llosera, L., Linares-Mustarós, S., Bikfalvi, A., & Coenders, G. (2020). A comparative assessment of graphic and 0–10 rating scales used to measure entrepreneurial competences. *Axioms, 9*(21). doi:10.3390/axioms9010021

Van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). Judgement rule aid for structural equation models version 3.0.4 beta [computer software].

Van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In A. Van Meurs. & W. E. Saris (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). Amsterdam: North Holland.

World Bank (2020). Individuals using the internet (% of population). Retrieved October 15, 2020, from https://data.worldbank.org/indicator/IT.NET.USER.ZS

Wu, H., & Leung, S. O. (2017). Can Likert Scales be treated as interval scales? A simulation study. *Journal of Social Service Research*, *43*(4), 527–532. doi:10.1080/01488376.2017.1329775

Yang, W., Moon, H. J., & Jeon, J. Y. (2019). Comparison of response scales as measures of indoor environmental perception in combined thermal and acoustic conditions. *Sustainability*, *11*(14), 3975. doi:10.3390/su11143975

Appendices

Appendix A:
Example smartphone screenshot

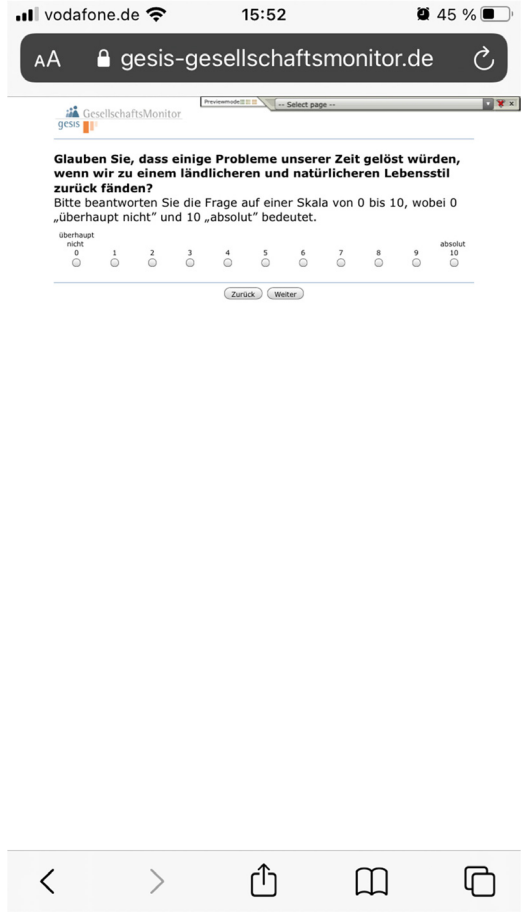


Figure 4 Depiction of the GESIS Panel web questionnaire on a smartphone: Showing the horizontal scale of method 3 (example: Experiment 1, trait 3)

Appendix B: Question formulations (original German versions)

Traits Experiment 1: Environmental attitudes

Trait 1	Können Sie sich mit Umweltschützern identifizieren?
Trait 2	Sollten wir alle bereit sein, unseren derzeitigen Lebensstandard zugunsten der Umwelt einzuschränken?
Trait 3	Glauben Sie, dass einige Probleme unserer Zeit gelöst würden, wenn wir zu einem ländlicheren und natürlicheren Lebensstil zurück fänden?

Traits Experiment 2: Supernatural beliefs

	Wie sehr glauben Sie an Folgendes?
Trait 1	An ein Leben nach dem Tod
Trait 2	An den Himmel
Trait 3	An Wunder

Appendix C:
Lisrel Input Base Model

```
! group 1
Data ng=3  ni=9 no=1368 ma=cm
km file=sb-group-1-corr.corr
mean file=sb-group-1-mean.mean
sd  file=sb-group-1-sd.sd
model ny=9 ne=9 nk=6  ly=fu,fi te=sy,fi ps=sy,fi be=fu,fi ga=fu,fi ph=sy,fi

! set lambdas of observed traits to 1, of not observed to 0
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
value 0 ly 7 7 ly 8 8 ly 9 9

! free error variances of all observed traits, set error variance of not observed to 1
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 1 te 7 7 te 8 8 te 9 9

! free trait gammas
fr ga 1 1 ga 2 2 ga 3 3 ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3

! set method gammas to 1
value 1  ga 2 4 ga 5 5 ga 8 6 ga 1 4  ga 4 5 ga 7 6
value 1 ga 3 4 ga 6 5 ga 9 6

! set trait variances to 1
value 1 ph 1 1 ph 2 2 ph 3 3

! free correlations among traits
fr ph 2 1 ph 3 1 ph 3 2

! free method variances
fr ph 4 4 ph 5 5 ph 6 6

pd
out mi iter= 5000 adm=off sc ec

! group 2
Data ni=9 no=1357  ma=cm
km file=sb-group-2-corr.corr
```

```
mean file=sb-group-2-mean.mean
sd  file=sb-group-2-sd.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in

! set lambdas of observed traits to 1, of not observed to 0
va 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9
value 0 ly 1 1 ly 2 2 ly 3 3

! free error variances of all observed traits, set error variance of not observed to 1
fr te 4 4 te 5 5 te 6 6 te 7 7 te 8 8 te 9 9
va 1 te 1 1 te 2 2 te 3 3

equal te 1 4 4 te 4 4
equal te 1 5 5 te 5 5
equal te 1 6 6 te 6 6

pd
out mi iter= 5000 adm=off sc ec

! group 3
Data ni=9 no=923 ma=cm
km file=sb-group-3-corr.corr
mean file=sb-group-3-mean.mean
sd  file=sb-group-3-sd.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in

fr te 1 1 te 2 2 te 3 3 te 7 7 te 8 8 te 9 9
va 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9 te 4 4 te 5 5
va 1 te 6 6

value 0 ly 4 4 ly 5 5 ly 6 6

equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
equal te 2 7 7 te 7 7
equal te 2 8 8 te 8 8
equal te 2 9 9 te 9 9

pd
out mi iter= 5000 adm=off sc ec
```

Appendix D:
Final Lisrel model adjustments, fit and JRule evaluation

Experiment	Mode	Model adjustments (in LISREL notation)	df	χ2	p-value	RMSEA	CFI	JRule
Environmental attitudes	Both	FR GA14 GA76 PH44(G3)	108	70.43	0,998	0,00	1,00	None
	Web	FR GA14 GA76 PH44(G3)	108	79.56	0,982	0,00	1,00	2
	Mail	FR GA14 GA45 GA76	108	78.95	0,984	0,00	1,00	6
Super natural beliefs	Both	FR TE66(G2)	110	297.19	0,000	0,04	0,99	4
	Web	VA 0 TE99(G3)*	112	231.61	0,000	0,04	0,99	2
	Mail	FR GA34	110	108.57	0,521	0,00	1,00	None

**Note*: When looking for a suitable model to analyse the answers of online respondents in Experiment 2, the best solutions found still resulted in a small negative error variance of the observed variable measuring trait 3 with method 3 (te 9 9), equal to -.01. However, given the fact that fixing this parameter to zero neither substantially affects the resulting estimates nor the fit of the model, we decided to accept the model with this parameter fixed to zero as our final solution in this case.

Appendix E:
Reliability, validity, and quality estimates for the different traits and methods for both experiments by mode

	Reliability				Validity				Quality			
	T1	T2	T3	Avg	T1	T2	T3	Avg	T1	T2	T3	Avg
Experiment: Environmental Attitudes												
Both modes												
M1 (Time 1)	.76	.76	.81	.77	.96	.90	.92	.93	.73	.68	.75	.72
M1 (Time 2)	.77	.79	.83	.80	.92	.76	.83	.84	.71	.60	.69	.67
M2	.81	.83	.86	.83	.86	.85	.86	.86	.70	.70	.75	.72
M3	.85	.83	.85	.84	1.00	.98	.98	.99	.85	.81	.83	.83
Web												
M1 (Time 1)	.77	.79	.83	.80	.98	.90	.94	.94	.76	.71	.78	.75
M1 (Time 2)	.77	.81	.85	.81	.94	.79	.86	.87	.73	.64	.73	.70
M2	.86	.85	.88	.86	.90	.86	.90	.89	.78	.73	.80	.77
M3	.86	.83	.85	.85	.98	.98	.98	.98	.85	.81	.83	.83
Mail												
M1	.77	.72	.77	.76	.86	.71	.77	.78	.67	.51	.60	.59
M2	.76	.81	.86	.81	.77	.83	.85	.82	.59	.67	.73	.66
M3	.86	.83	.83	.84	1.00	.90	.90	.94	.86	.75	.75	.79
Experiment: Supernatural Beliefs												
Both modes												
M1	.96	.96	.94	.95	.94	.94	.94	.94	.90	.90	.89	.90
M2 (Time 1)	.96	.94	.94	.95	.96	.94	.94	.95	.92	.89	.89	.90
M2 (Time 2)	.96	.94	.88	.93	.96	.94	.94	.95	.92	.89	.83	.88
M3	.98	1.00	.98	.99	1.00	1.00	1.00	1.00	.98	1.00	.98	.99
Web												
M1	.96	.96	.92	.95	.94	.94	.94	.94	.90	.90	.87	.89
M2	.96	.94	.94	.95	.96	.94	.94	.95	.92	.89	.89	.90
M3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mail												
M1	.98	.96	.90	.95	.94	.92	.94	.93	.92	.89	.85	.89
M2	.94	.94	.88	.92	.94	.92	.92	.93	.89	.87	.81	.86
M3	.98	1.00	1.00	.99	.98	.98	.98	.98	.96	.98	.98	.97

Note: M=Method, T=Trait, M1=7-point fully labelled horizontal, no scale definition; M2=101-point numerical open-ended scale, scale definition present; M3=11-point partially labelled horizontal, scale definition present; Avg=Average.